卒業論文

日本語文中の英文字の認識手法の開発

東北大学工学部情報工学科４年
中川 修司

目 次

図目次

表目次

第1章

序論

1.1 本研究の背景

計算機の発達、普及に伴い、人間の能力を機械(計算機)によって実現しようという努力が現在も行われている。中でも、人間の文字を読み取るという能力の実現である文字認識技術については古くから盛んに研究が行われている。その理由の1つとしては、文字が1文字に1つの概念が対応し、記録性もよく、我々にとって身近なものであるため、研究対象になりやすいことが挙げられる。また、もう1つの理由としは、文字を読み取る技術が計算機へのデータ入力の主流であるキーボードを介した作業を省力化、効率化するための装置の開発に結びつき、社会的なニーズも強かったことが挙げられる。

当初、光学文字読取装置(Optical Character Reader, 略してOCR)は郵便番号の自動読取区分装置に代表されるような、字種(英数字記号のみ、片仮名も含む、漢字・平仮名も含む等)、変形自由度(活字、手書き等)、筆記具(鉛筆、OCRボールペン、無制限等)、字体(単一、複数等)、文字ピッチ(固定、自由等)、搬送系(OCR用紙、上質紙、普通紙等)、識別水準(個別、単語、文章等)、フォーマット制御(固定、自由等)に制約がある文書に対してのものが実用化されてきた。しかし、ワードプロセッサ、パーソナルコンピュータの普及によって、多様な字種、字体、ピッチの文書の作成が容易になり、また、文書のフォーマットも複雑化、多様化してきたこたにより、制約の少ない(無い)OCRの開発が望まれてきた。

最近では、漢字認識、手書き文字認識の発展、半導体技術の進歩によって、これまで困難とされてきた複雑かつ高度な処理が可能なものになってきている。また、ワークステーション、パーソナルコンピュータの普及によって、膨大な量の文書(新聞、雑誌、書籍等)のデータベース化の社会的要望が高まっている。このため、より制約の少ないOCRの開発、実用化の努力がなされ、OA(オフィスオートメーション)、FA(ファクトリーオートメーション)、EA(エンジニアオートメーション)等の各分野で応用されてきている状況である。

このように、文字認識技術はOCRの開発、実用化という大きな成果を挙げている。しかし、現在のOCRはどのような文書でも認識できるわけではないし、文字認識能力も人間に比べるとまだ劣っている。よって、より完璧なOCRの開発、実用化が求められてい

る [1][2]。

1.2 本研究の目的

文字認識はいくつかのステップを通して行われるが、中でも重要であるのが文字切り出し技術である。本研究では、文字切り出しを高精度に行わせるための前処理として、英文字抽出法を提案する。

これまで、日本語文書用の文字切り出し法や英語文書用の文字切り出し法は数多く提案されているが、英文字の混在している日本語文書に対してはあまり提案されていない。なぜならば、英文字が混ざった日本文は英文字の不定なピッチと大きさによる切り出しの難しさと、日本語の分離文字による切り出しの難しさが互いに影響し合い、文字切り出しをより難しくしているからだ。

この問題を解消するための1つの方法として、今回、あらかじめ英文字部分を抽出することによって、従来提案されている日本語文書用の文字切り出しと英語文書用の文字切り出しを併用してより高精度な文字切り出しを実現しようというわけである。そして、文字認識全体としての認識率の向上を図るのである。

1.3 本論文の構成

本論文の構成は以下の通りである。

第1章 序論であり、本研究の背景と目的について述べる。

第2章 これまで提案されてきた文字切り出し法とその原理をまとめ、そこに含まれている問題点を明らかにする。

第3章 第2章で述べた問題点をふまえ、ここに新たに文字切り出しの前処理として、英文字抽出法を提案する。

第4章 第3章で提案した英文字抽出法の効果を認識実験を通して検討する。また、このために、第2章の問題点を解消するための簡易的な文字切り出し部、辞書を作成する。

第5章 本研究の結論、今後の課題を述べる。

第 2 章

文字切り出しとその問題点

2.1 はじめに

一般に文字認識は観測、前処理・正規化、特徴抽出、識別の 4 つ過程によって行われている。まず、観測とは、用紙から文書画像をスキャナ等で読み取ることである。次に、前処理・正規化とは、読み取った文書画像を文字単位に分割し、特徴抽出し易いように変形することである。そして、特徴抽出とは、認識のための情報を文字毎に抽出することである。最後に、識別とは、文字毎に得られた特徴と標準パターンを参照して各文字を断定することである。完全な文字認識のためには、これら 4 つの技術の発展はどれも必要不可決であるが、本研究では、中でも前処理部分に着目する。

前処理にもいくつかの段階 (スムージング、文字切り出し、正規化、細線化等) があるが、中でも文字切り出しは重要な技術と言える。なぜならば、文書画像から文字を正確に抽出できなければ、どんなに素晴らしい特徴抽出部、識別部があっても役に立たないものになってしまうからである。

これまで、いろいろな文字切り出し法が多くの研究者たちによって提案されてきているが、全ての文字に対して有効な方法はまだ発見されていない。ここでは、これまでの文字切り出し技術を振り返り、それらに残されている問題点をまとめることにより、新たな文字切り出し法の可能性を考える。

2.2 基本的な文字切り出し法

初期の OCR では文字の大きさ、配置が決まっている文書 (帳表等) を対象にしていたため、文字の切り出しの必要は無かった。しかし、OCR 技術の発展には、大きさ、配置等の制約の無い文書への応用が必要であり、それを実現するためには文字切り出し技術が重要となってきた。このため、様々な文字切り出し法がこれまで考案されてきている。

2.2.1 射影 (投影) 分布を用いる方法

射影 (投影) 分布とは、与えられた画像をある方向に投影した時にあらわれる黒画素の分布のことである。具体的には、投影方向に沿って計測した画素の度数分布のことである。

これを用いた切り出し方法 [3] は次のように行う。あらかじめ行単位に分離された画像に対して垂直方向に射影 (投影) 分布をとり、分布の山の部分に文字らしきものがあると判断し、分布の谷 (画素数 0) の部分を境界にして切り出していく。

この方法では、最初から垂直方向に各文字が分離していなければ、切り出しミスが生じてしまう。

図 2.1: 射影 (投影) 分布を用いた文字切り出しの例

2.2.2 連結領域に着目する方法

黒画素が連結している領域を文字らしいとして、連結領域毎に切り出していく方法 [4] である。連結の概念には 4 連結、8 連結 [5] があるが、文字の形状特徴 (縦線、横線以外に斜め線、曲線もありうる) を考慮して、主に 8 連結が用いられている。

この方法では、垂直方向に各文字が分離していなくても連結さえしていれば切り出しは成功する。しかし、文字自身が最初から分離している場合は切り出しミスになってしまう。

2.2.3 その他

前出の 2 つの方法が広く用いられているのであるが、その他にも切り出し方法は提案されている。例えば、統計的なモデルで文字ピッチを推定して、動的計画法を用いて切り出しを行う方法 [6] が考案されている。この方法は、固定ピッチの文書では効果的だが、欧文を含むような不定ピッチには適用できないという問題点がある。

2.3 基本的な文字切り出し法の問題点

日本語の特徴、英語の特徴等により、前節で紹介した切り出し方法では完全に個々の文字を切り出すことは難しい。以下に日本語、英語の各々の場合の問題点を挙げてみることにする。

2.3.1 文字切り出しにおける日本語の特徴上の問題点

日本語のみの文字切り出しには以下のような問題点がある。

1. 分離文字の存在。

図 2.2: 分離文字の例 (「に」が分離している)

2. つぶれ等による文字同士の接触。

図 2.3: 文字の接触の例 (「組織」が接触している)

2.3.2 文字切り出しにおける英語の特徴上の問題点

英文字の切り出しのみを考えると、以下のような問題点が挙げられる。

1. 不定な文字ピッチ (3.4.1 参照)

   l と m、i と w のように、文字間のピッチに大きな開きが見られる。

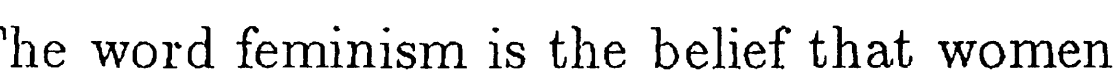


図 2.4: 英語の一般文書 (不定ピッチ) の例

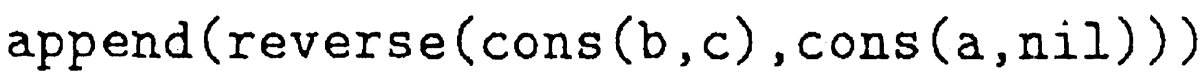


図 2.5: 英語の一般文書 (固定ピッチ) の例

2. 入り組み文字

- 印刷後の仕上がりを美しく見せるための文字間隔を詰める処理 (kerning)

図 2.6: kerning の例 (“aj” が入り組んでいる)

- 文字自体が斜めに傾いている斜体 (イタリック体等)

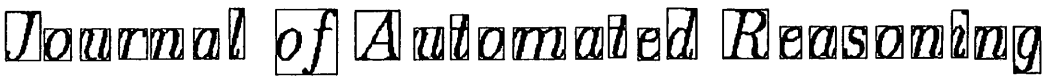


図 2.7: イタリック体の例 (“of” が入り組んでいる)

3. 文字間の接触

- 印刷の際、決まって 2 つ以上の文字が接触してしまう連字 (ligature)

図 2.8: ligature の例 (“ffi” は f と f と i が接触してできた連字)

- 人間が書く筆記体のようにわざと接触させて書れた文字
- コピーなどの結果、つぶれが生じてしまい、接触してしまった文字

2.3.3　対策

このように、日本語には日本語の、英語には英語の切り出しにおける問題点がある。よって、日本語 OCR では日本語を正確に切り出すための処理 (ピッチ情報を用いた分離文字の統合等) を、英語 OCR では英語を正確に切り出すための処理 (接触文字の分離 [7]、文字単位の傾き補正 [7] 等) をしている。

2.4 日本語と英語が混在している文書の切り出し法

一般に日本語文書は漢字、平仮名、片仮名の他に英数字を用いていることが多く、英数字の大きさ、配置の不規則さによる生じる問題と日本語の分離文字によって生じる問題とが互いに影響し合い、文字の切り出しを更に難しくしている。

これまで、この問題を解決するため、ピッチ情報を用いて切り出す方法 [8]、文字列のレイアウト上の特徴を用いて切り出す方法 [9]、容易に切り出せる文字から切り出していく方法 [10] 等、様々な手法が提案されてきている。

2.5 まとめ

このように、現在の文字切り出し技術では、日本語のみの文書に対して、英語のみの文書に対して、日本語と英語が混在していても定ピッチな文書に対しての3つの場合はほぼ解決したと言える。しかし、日本語と英語が混在した不定ピッチの文書に対する効果的な切り出し法はあまり考案されていない。よって、本研究では、日本語と英語が混在した文書の文字切り出しのために、文書全体から英語の領域を抽出することを目標とする。そうすることによって、英語の領域は英語用の切り出しを、その他の領域は日本語の切り出しをすることができるようになり、全体的な切り出し成功率が高まるだろうと考えられる。英語の抽出法は次章で提案する。

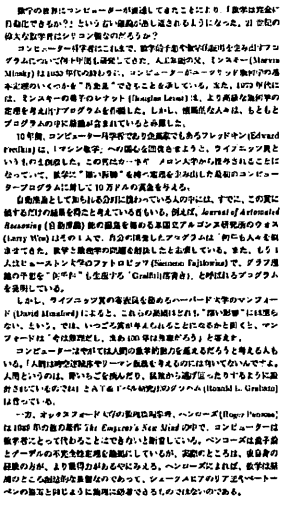
第 3 章

日本語文中からの英文字抽出

3.1 はじめに

日本語文書の中には、種類によって英文字が混ざっていることがある。例えば、科学技術関連の文書 (論文等) がその代表である。日本語と混ざっている英文字は文章内の文字ピッチを乱すため、文字の切り出しが難しくなり、その結果、認識率を低下させる原因となっている。

この章ではこの文字ピッチの乱れを逆に利用して、文章中から英文字部分の抽出をする。図 3.2 のフローチャートは英文字抽出の一連の流れを示したものである。次節からは、それぞれの段階ごとに説明していくことにする。



数学の世界にコンピューターが進出してきたことにより、「数学は完全に自動化できるか?」という古い議論が蒸し返されるようになった。21 世紀の偉大な数学者にシリコン製なのだろうか?

コンピューター科学者にこれまで、数学命題を証明や反証(証明を含み出すプログラムについて何十年にも研究してきた。人工知能の父、ミンスキー(Marvin Minsky) は 1960 年代に早くも、コンピューターがユークリッド幾何学の基本定理のいくつかを"自発見"できることを示している。また、1970 年代には、ミンスキーの弟子のレナット(Douglas Lenat) は、より高度な数学の定理を考え出すプログラムを作成した。しかし、楽観的な人々は、もともとプログラムの中に答えが含まれていたと非難した。

10 年前、コンピューター科学者であり企業家でもあるフレッドキン(Edward Fredkin) は、「マシン数学」への関心を回復させようと、ライプニッツ賞という賞を創設した。この賞は、カーネギーメロン大学から授与されることになっていて、数学に"賢い振舞"を持つ定理を証明した最初のコンピュータープログラムに対して 10 万ドルの賞金を与える。

自動推論として知られる分野に関わっている人の中には、すでに、この賞に値するだけの結果を得たと考えているものもいる。例えば、*Journal of Automated Reasoning* (自動推論) 誌の編集を務める米国立アルゴンヌ研究所のウォス (Larry Wos) はその 1 人で、自分の開発したプログラムは「約二十人々を悩ませてきた、数学と論理学の問題を解決したことを証明している。また、もう 1 人は ヒューストン大学のファトロピッツ(Nicholas Fatlopirra) で、グラフ理論の予想を"反例"も生成する「Graffiti(落書き)」と呼ばれるプログラムを開発している。

しかし、ライプニッツ賞の審査員を務めるハーバード大学のマンフォード (David Mumford) によると、これらの結果は「賢い振舞」には該当しない、という。では、いったい何が考えられることになるかと聞くと、マンフォードは「それは原理的に、あと 100 年は有望だろう」と答えます。

コンピューターはやがては人間の数学的能力を超えるだろうと考える人もいる。「人間は時空連続体やリーマン幾何を考えるのには向いていないんですよ。人間というのは、赤いものを飲んだり、猛獣から逃げ回ったりするように設計されているものですよ」とAT&T ベル研究所のグラハム (Ronald L. Graham) は言っている。

一方、オックスフォード大学の数理物理学者、ペンローズ(Roger Penrose) は 1989 年の著書 *The Emperor's New Mind* の中で、コンピューターは数学者にとって代わることはできないと断言している。ペンローズは量子論とゲーデルの不完全性定理を融合しているが、実際のところは、直感の鋭敏の方が、より説得力があるように思える。ペンローズによれば、数学は最高のところで創造的な営みなのであって、シェークスピアやリア王やベートーベンの創造と同じように機械に必要できるものではないのである。

図 3.1: 英文字が混在する日本語文書例

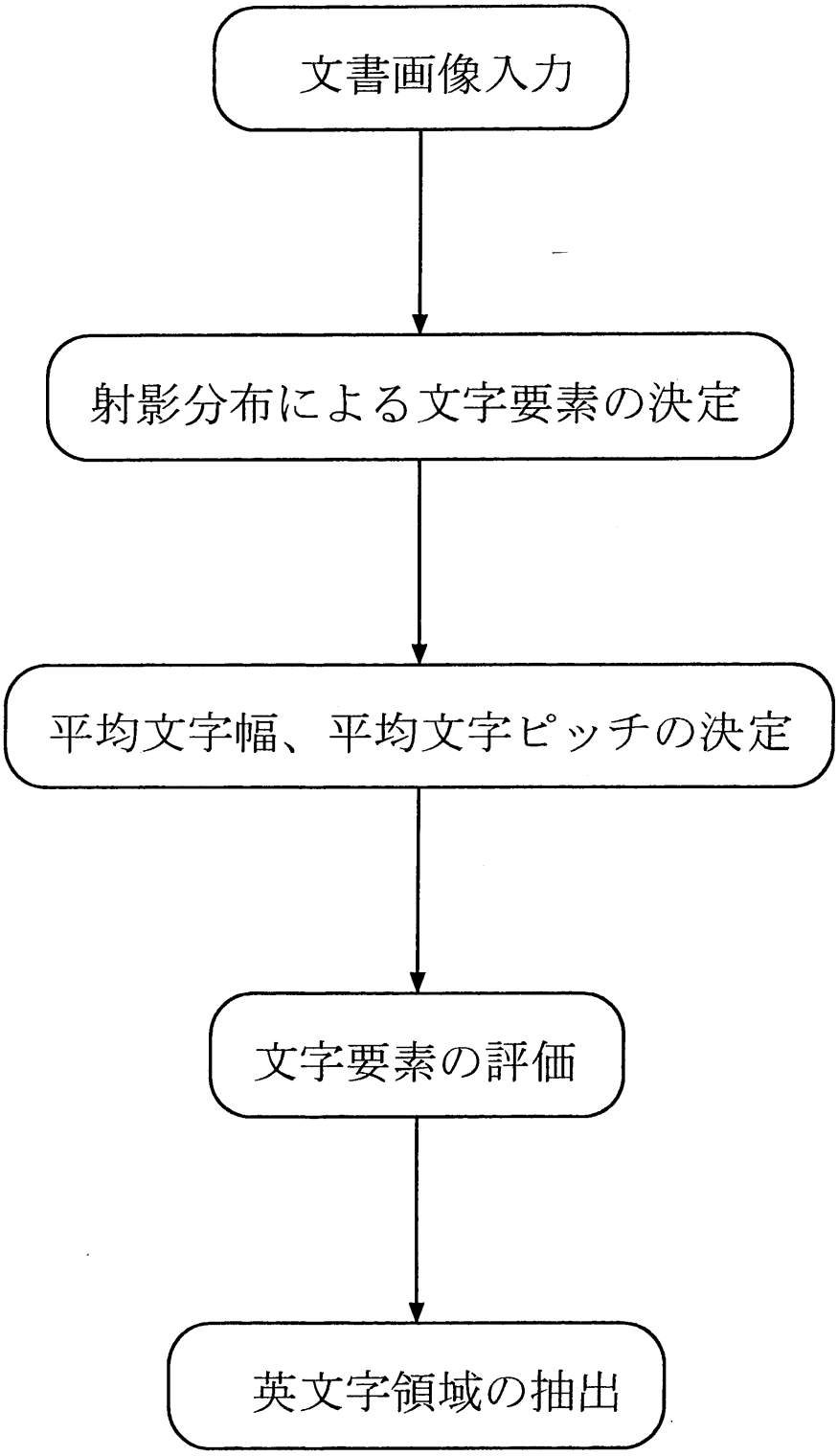


図 3.2: 英文字抽出の概略

3.2　文書画像入力

認識対象となる文書は 1 ページずつスキャナで読み込む。この時の読み取り密度は 400dpi である。そうして得られた文書画像に対して、次節以降の処理を施すのである。

本研究では横書きの日本語文書 (特に英文字を含んだ) を中心に扱っていく。

3.3　射影分布による文字要素の決定

本研究では文書画像から行を抽出し、その行から各文字要素を決定するために射影分布 (2.2.1 参照) を利用している。以下にその方法を簡単に説明する。

まず、1 ページ分の文書画像が与えられたならば、その文書に対して水平方向に射影分布をとる。そして、求めた射影分布の山の部分に行らしきものがあると判断して、射影分布の谷の部分 (画素数 0) を境界として行を抽出するのである (図 3.3 上)。

次に行の画像が得られたならば、各行ごとに文書に対して垂直方向に射影分布をとる (図 3.3 中)。そして、行抽出と同様に射影分布の山の部分を文字らしきものがあると判断し、それを文字要素として登録する (図 3.3 下)。ここで言う登録とは、その文字要素の番号 $i \in [0, num)$(その行に i 番目に登場) を与え、その高さ height(i) と横幅 width(i)、そして、横方向の座標 s-width(i) の各データを保存することである。これらのデータは次節以降に用いることになる。

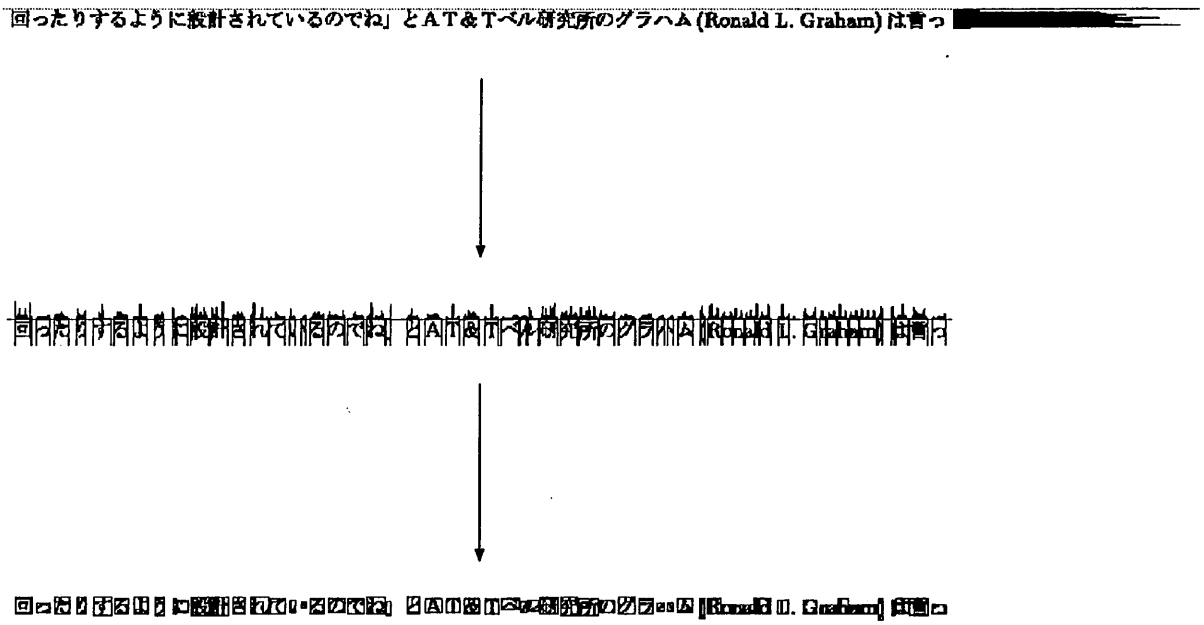


図 3.3: 文字要素決定までの過程

3.4　平均文字幅、平均ピッチの決定

日本語文書を行単位で見ると、日本語(漢字、平仮名、片仮名、記号類)のみしかあらわれないものと日本語に加えて英文字もあらわれるものと2種類に分類できる。(英文字と記号類のみしかあらわれないものがある場合も考えられるが、それは稀であるので分類には含んでいない)。それらの中で日本語と英語との間の何らかの違いを見出すことができれば英文字抽出も可能になってくる。この節では、その違いを文字の外見上の特徴のみ(本研究では文字幅と文字ピッチ)で発見し、それを英文字抽出の方法に取り入れることを考える。

3.4.1　文字幅と文字ピッチ

文字幅とは単純に文字の横幅のことである。ただし、本研究ではこの時点で完全に個々の文字を切り出してはいないので正確には文字要素の横幅を指している。また、文字ピッチとは隣合う文字(要素)とその文字(要素)自身との間にできる空白の半分をその文字(要素)の横幅に加えたものである。これらを式であらわすと以下のようになる。

$$\text{文字幅} \ : \ \mathrm{Width}(i) = \mathrm{width}(i)$$

$$\text{文字ピッチ} \ : \ \mathrm{Pitch}(i) = \mathrm{width}(i) + \mathrm{space}(i) + \mathrm{space}(i+1)$$

但し、

$$\mathrm{space}(i) = \begin{cases} \frac{\text{s-width}(i+1) - \text{s-width}(i) - \text{width}(i)}{2} & (0 < i < num) \\ 0 & (i = 0, i = num) \end{cases}$$

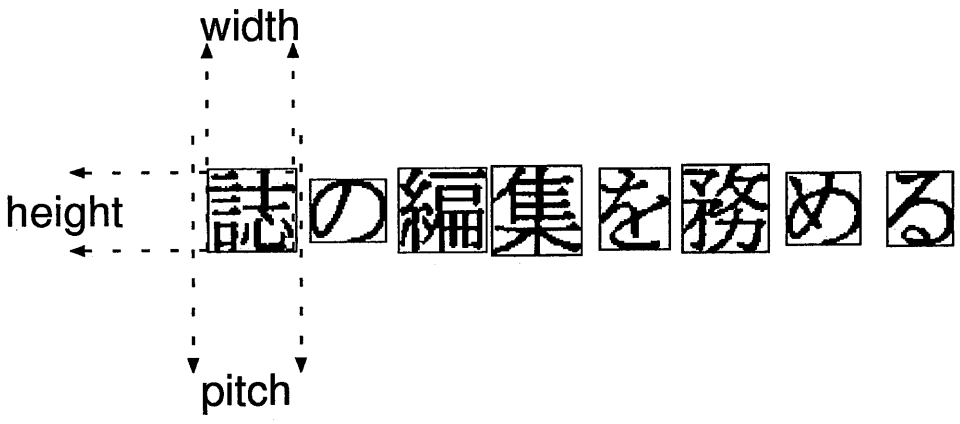


図 3.4: 文字要素の外見情報の定義